

AD 694590

THE THESAURUS IN ACTION

Background Information  
for a Thesaurus Workshop  
at the  
32nd Annual Convention of the

AMERICAN SOCIETY FOR INFORMATION SCIENCE

October 1969  
San Francisco, California

Prepared by the Workshop Panel



Reproduced by the  
CLEARINGHOUSE  
for Federal Scientific & Technical  
Information Springfield Va. 22151

has been approved  
for reproduction and sale; its  
distribution is unlimited.

42

Security Classification

## DOCUMENT CONTROL DATA - R &amp; D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Information Systems Office Office Chief of Research & Development Department of the Army, Washington, D. C. 20310		2a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	
		2b. GROUP	
3. REPORT TITLE THE THESAURUS IN ACTION: Background Information for a Thesaurus Workshop at the 32nd Annual Convention of the American Society for Information Science, October 1969, San Francisco, California			
4. DESCRIPTIVE NOTES (Type of report and inclusive dates) Panel members' presentations			
5. AUTHOR(S) (First name, middle initial, last name) Pamely C. Daniels, Chairman. Department of the Army, CRDISO. Terry L. Gillum and William Hammond, Automated Systems Corp. James G. Peirce, Frankford Arsenal. Frank Y. Speight, Engineers Joint Council			
6. REPORT DATE 1 October 1969	7a. TOTAL NO. OF PAGES 33	7b. NO. OF REFS 5	
8a. CONTRACT OR GRANT NO.	8a. ORIGINATOR'S REPORT NUMBER(S)		
b. PROJECT NO.			
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)		
d.			
10. DISTRIBUTION STATEMENT This document has been approved for public release; its distribution is unlimited.			
11. SUPPLEMENTARY NOTES Thesaurus Workshop sponsored by American Society for Information Science, Charles P. Bourne, Pres. Elect & Convention Chm.		12. SPONSORING MILITARY ACTIVITY	
13. ABSTRACT Five facets of background information are intended to be helpful to persons interested in the construction and use of a thesaurus, in the sense of a standardized language for communication with and between information storage and retrieval systems. Based on TEST as an example. First, the thesaurus is described and defined, both as to what it is and what it is not. Then, the story of Project LEX shows how a thesaurus is developed, from rules and conventions to published format. Next, the building of a local specialized thesaurus compatible with the broader standard is explained. Then, application of the thesaurus to systems of indexing other than documents is described. Finally, experience with the thesaurus is discussed, and guidance outlined on updating and improvement. The authors are the members of the ASIS Workshop Panel.			
KEYWORDS: Thesauri, Index terms, Communication standards, Thesaurus construction			

DD FORM 1473  
1 NOV 65REPLACES DD FORM 1473, 1 JAN 64, WHICH IS  
OBSOLETE FOR ARMY USE.

Security Classification

## THE THESAURUS IN ACTION

### Contents

Identifying Information . . . . .	3
Parmely C. Daniels	
Comments on the TEST Conventions . . . . .	7
Terry L. Gillum	
Satellite Thesaurus Construction . . . . .	14
William Hammond	
Use of TEST in the Preparation of the Research & Development Capability Index . . . . .	26
James G. Peirce	
Plans for Updating TEST, and Survey of Users . . . . .	33
Frank Y. Speight	

## THE THESAURUS IN ACTION

### Preface

The papers presented here are intended to be helpful to persons interested in the construction and use of a thesaurus, in the sense of a standardized language for communication with and between information storage and retrieval systems. Many people are hesitant to get involved with a thesaurus, partly because there is scant clear guidance, partly perhaps because change of system is feared to mean total disruption and conversion of a going system, and partly because a quick look at a handy thesaurus shows that it doesn't fit the local situation. These hurdles hopefully are lowered or removed.

There are many thesauri in existence, admirably suited to their purpose. Simply for consistency of illustration and example, these papers are concerned with the Thesaurus of Engineering and Scientific Terms (TEST), as developed under Project LEX, jointly by the Defense agencies and the Engineers Joint Council. Also, for illustration of the use of the thesaurus, the Termatrix system of visual coincidence indexing is used.

The papers present five facets of background information. First, the thesaurus is described and defined, both as to what it is and what it is not. Then, the story of LEX shows how a thesaurus is developed, from rules and conventions to published format. Next, the building of a local specialized thesaurus compatible with the broader standard is explained. Then, application of the thesaurus to systems of indexing other than documents is described. Finally, experience with the thesaurus is discussed, and guidance outlined on updating and improvement.

The authors of these papers are members of the ASIS Thesaurus Workshop Panel. They are briefly identified as follows, in the order in which their papers appear.

Parmely C. Daniels, Panel Chairman. Advanced Information Techniques, Department of the Army, Office Chief of Research & Development, Information Systems Office. R&D Member, Advisory Panel, Dept./Army Vocabulary of Information Elements (DAVIE); Army Member, Project LEX Steering Group; instrumental in original Army STINFO Program. Developed and monitored Project ACSI-MATIC for Army Intelligence Visiting Professor, University of Puerto Rico, and Consultant to Government of Puerto Rico during Operation Bootstrap. Personnel Officer of various agencies. Member of "Hawthorne Experiments." Drake University (AB Astron), University of Iowa (MA Psychol), Harvard Business School. ASIS

Terry L. Gillum. Automated Systems Corporation. Recently, Data & Information Management Systems Dept., Military Systems Division, System Development Corp., where he helped compile a special applications thesaurus within Dept. of Defense. With DDC (and ASTIA), as indexer, abstracter, searcher and lexicographer,

participated in three editions of DDC Thesaurus. In various capacities, assisted with TEST Rules & Conventions and compilation of the Thesaurus. Served on Engineering Vocabulary Panel, and Panel on Educational Terminology. Also on COSATI Subpanel on Classification and Indexing. Oklahoma State University (AB). ASIS.

William Hammond. Manager, Education System Division, Automation Systems Corporation. Past 11 years, manager or principal investigator on major projects of design, implementation and operation of automated technical information systems, and compatibility among the large federal information systems. Designed and produced software packages to support compilation and maintenance of thesauri from the first ASTIA Thesaurus, and including TEST, NASA Thesaurus, Water Resources Thesaurus (for SIE), Urban Thesaurus (for Kent State University), Behavioral Thesaurus (for NAS), US Army Biological Labs Thesaurus, S. C. Johnson (Johnson's Wax) Thesaurus, Linguistic Thesaurus (for Center for Applied Linguistics), and Department of State Thesaurus. Duke University (Civil Engineering), American University (Pub Adm). ASIS.

James G. Peirce. Technical Information Officer, Action Officer for Army QRI (Qualitative Requirements Information) Data Files Program, Frankford Arsenal, Pa. Fifteen years in printing and newspaper publishing business. Materials Engineer, Ordnance Engineer, Publications Writer, Editor, Administrative Officer, and Program Planning Officer, Frankford Arsenal. In scientific and technical information program and QRI since 1958. Haverford College (chemistry), University of Colorado (AB 1938). ACS, AAAS, RESA, ASIS.

Frank Y. Speight. Director Information Program, Engineers Joint Council since 1964; Secretary & Program Manager for The Tripartite Committee since 1967. Supervised publication of EJC's first thesaurus and directed EJC's effort in cooperation with Project LEX to produce the TEST. With EJC and The Tripartite Committee, directed programs leading to improved information systems in engineering. BS in chemical engineering, specialized in engineering materials in research, advisory services, and standardization. ASIS, ASTM, ACS, and Fellow of AAAS and AIC.

#### Acknowledgments

Copies of the TEST for use by Workshop participants are loaned through courtesy of the Engineers Joint Council.

Termatrix demonstration kits are furnished to the Workshop by courtesy of the Jonker Corporation.

## IDENTIFYING INFORMATION

Parmely C. Daniels

Department of the Army  
Research & Development Information Systems Office

This introductory section proposes a common understanding as to the concept of a thesaurus as applied today to information storage and retrieval in science and technology. Perhaps arbitrary, but hopefully acceptable description of common ground that is not yet well defined sets the stage for the sections to follow. Discussed are, what a thesaurus is, what it is not, and where and how it is used.

### What a Thesaurus Is

Verbal communication between two people is the attempt to convey meaning by words. To the extent that the words mean the same to both parties, there is understanding. Understanding depends upon common meaning of language.

When communication is oral between two persons, they can haggle with questions and definitions and synonyms and illustrations until they arrive at common meaning.

When a person is reading text, he is more dependent upon knowledge of meanings in advance. He can use a dictionary or he can read on, in the hope that the context or expansion will help, but the text can't tell him more than is on the page.

When he addresses a machine, he is even more restricted. Devoid of prose, paraphrase and restatement, the machine means exactly what someone has made it mean. Until the machine is instructed in the user's meanings, or vice versa, that is, until there is an agreed standard meaning between machine and users, communication is unsatisfactory. This communication link of agreed meanings is a thesaurus, whether programmed into the machine, hung over the console, or published in a book.

Communicating with machine retrieval systems is a growing way of life, whether at the television, in the automat, on the jukebox, at the pushbutton telephone, or from a more sophisticated machine such as a microform selector or a computer data bank. People do sometimes make mistakes, and sometimes the bin behind a label may have nothing in it, but when a considered selection of a retrieval term has been made and the machine emits a wrong product, one of two judgments has been in error. Either the wrong product was put in the bin or the wrong term was selected. Both errors are due to a difference in understanding between the person who put the material in and the person who took it out as to just what the label meant. When the labels are many and overlapping or vague, an

agreed definition must be established as to just what goes under each label. This set of labels and their discrete scopes, when agreed to by the community involved, is a thesaurus.

Even when words are assumed to be understood, there can be inadequate or frustrating communication. One person asks about cloth, another about fabric, a third about textiles. Do they all want the same product? Only agreed definitions will tell. Without them, one would have to remember to try all three labels, and maybe twill, satin and damask in addition, to uncover what he really seeks. Or, of course, it is possible to instruct the machine to identify all these for him. There is a wide range of attitudes based on these alternatives.

#### Where it is Used

At one extreme of these "schools of thought" are those who insist that in talking to a system no one should use a word that is not "standardized," because any other word can be misunderstood. These people must be sure that every word they use is in the thesarus before they try it on the system or else the system may turn back garbage or a blank because it doesn't know the meaning of the terms. At this extreme, every author and every searcher needs to keep a thesaurus and use it like a telephone directory every time he wants to submit information or call for it. To this group, the idiot machine which responds to their will can do no wrong. Failures are the fault of the user.

At the other extreme are the "freedom-of-speech" people who insist that anyone who is forced to say exactly what he means in terms no one can misunderstand is thereby unduly handicapped, and he may lose his vague idea altogether by trying to make it precise and meaningful to an idiot machine. For him, the user is the only one who is right, because he is the only one who knows what he wants. Any faults of delivery are neither with him or with the machine, but with the stupid programmer who didn't tell the system how to interpret his rhetoric. Instead of rhetoric, these people say "natural language." Here the thesaurus is programmed into the machine, and only the programmer has the benefit of knowing the language. There are no copies for customers or contributors to be helpful to the system or to devise compatible vocabulary for communication among themselves.

But, like the six blind men who spoke so wisely about the elephant but none identified it, this range of people are talking about a thesaurus without identifying it. At some point, in any communications concept, there must be an interface between a mind and a machine where the meaning in mind is put into symbols that the idiot machine or stock boy understands and can respond to. This can only be done with standard unique meanings for terms, and this is a thesaurus.

Practically, the interface is distributed between the two extremes. An author can help the validity of the information system if he will try to express his meaning for indexing purposes in standardized terms. If he does not, a documentalist must guess the appropriate near-synonyms in the thesaurus as best he can in order to index it for him. There is usually some combination of effort, in which the

author or user is provided access to the thesaurus, but is encouraged to use any language which will better describe the contents from the user's point of view. This process continually validates and updates the thesaurus.

### Keywords versus Data Elements

A rather vague distinction is maintained between thesaurus keywords and management data elements and items. Both are standardized terms identifying discrete or unique scopes of information for labeling purposes. A keyword usually identifies some substantive content or subject matter of a document -- that is, what the document is about. It identifies information, rather than an object. A keyword may represent one element of a broader keyword, or in turn may comprise a number of narrower terms, but they all refer to a scope of subject-matter content of information. A data element or data item, on the other hand, is the name of a person, place, or thing, or a class of persons, places, or things. The difference between keywords and data elements or items is not that clear, however, for the data elements are seldom used to identify the persons, places or things themselves, but documents about or by them, or records of them, or addresses where more information can be found. Keywords and data elements are found together in the same data bank and thesaurus, because documents are located not just by subject matter content, but by author, date, title, project number, sponsor, and other data elements. Similarly, infrared detectors and dielectric lenses can be subject matter of reports, or be supply items. Actually, the principal difference is that the technical information people and the management information people got started independently and have never developed a close working relationship.

### What a Thesaurus Is Not

#### Not a Dictionary

A dictionary is used, among other things, as a means of finding the meaning or meanings for a word, while a thesaurus is intended as a means of finding the unique word for a given meaning. A thesaurus is more like a glossary, in that it is made up of terms important or peculiar to a field of knowledge, rather than like a dictionary which includes total vocabulary and all parts of speech.

#### Not an Index

The terms in a thesaurus may be used, in whole or in part, as an index, but the index is the system of terms and their codes actually used in a collection, while the thesaurus identifies the scope and uniqueness of every standardized term, and its relationship to other terms, whether the terms are used in a given system or not. The thesaurus is the arbiter -- the authority -- the standard -- on which to establish compatibility and communication among a community of index systems. So, it tells what the terms in an index are about, but it is essentially a reference rather than a working document.



### Not a Classification System

While a thesaurus does delimit the scope of terms by broader and narrower terms, it is not done for the purpose of orderly arrangement on a shelf, or an area in which to browse, or the unique place to store a book. In fact, a thesaurus vocabulary is most effective for storage and retrieval when unrelated keywords are chosen to identify a document. For example; Seashells, Collecting, Caribbean, Guides, is a much clearer identification than Zoology, Invertebrates, Mollusca, Popular Works.

### Information by Coincidence

In most of today's collections, a document is given an "address," which need not have any other meaning, and all clues to the document's identity are keyed to that address. When the same address responds to a number of clues or tests, the combined descriptors can produce a real pinpointing. In fact, thoughtful configurations may produce the stimulus to new knowledge or new applications of knowledge

The machine systems of searching on combinations of keywords are closely parallel to manual systems of inverted indexing, or coordinate indexing, or visual coincidence indexing, as you may have heard it. It is called inverted because, instead of a card representing a document with all the keywords on the card and usually a copy filed under each keyword, the card represents a keyword, and all the documents to which it applies are identified on it. This keyword card thus becomes the identifier of a bibliography on the subject of the keyword. It is called coordinate because, when two or more cards are compared the duplications of address represent coordinated coverage of the combination. Now, if the documents are referenced by punched holes in dedicated spaces, their superposition will show up the coincidences visually. This brief explanation is made because it is practical to illustrate the use of a thesaurus with examples using a manual coordinate system.

### Definition

In summary, for purposes of this exercise, the following definition of a thesaurus is proposed:

**Thesaurus:** An organized reference of the terms accepted and approved as a standard by participating members of a specialized population in a defined area of information, which identifies the scope of each term by inclusions, exclusions and associations, so that all terms are clear and discrete and in the aggregate are comprehensive for communication and identification of information in the defined area.

## COMMENTS ON THE TEST CONVENTIONS

Terry L. Gillum

System Development Corporation

### Introduction

An important part of developing the Thesaurus of Engineering and Scientific Terms (TEST) was the documentation of the guiding principles under which it was to be constructed. These principles or conventions were published before work on the thesaurus was begun and have since been reprinted by Engineers Joint Council (EJC) as Thesaurus Rules and Conventions. This statement of conventions has generated considerable interest -- and perhaps some confusion and controversy -- among those interested in thesauri. The purpose of this presentation is to explain the rationale of the conventions and to discuss their application in the development of TEST.

### Background

TEST was originally conceived as a model and base for the development of subject indexing vocabularies within the Department of Defense, just as the Thesaurus of Engineering Terms of EJC Thesaurus had been intended to be used as a model within some parts of the engineering community. It was largely a matter of coincidence that a revision of the EJC thesaurus was planned for about the same time that the DOD thesaurus was to be compiled. The timing and the similar purposes of the two efforts lead to the joint DOD-EJC undertaking that was called Project LEX.

In both the EJC and the DOD efforts, the importance of establishing guidelines before the actual thesaurus work began had been recognized. The DOD project was charged with developing a detailed statement of thesaurus conventions as its initial product. A committee organized by EJC had been at work for several months on a statement of "Rules for Preparing and Updating Engineering Thesauri." One of the first orders of business of the joint effort was to produce a consolidated statement of the guidelines as part of the Manual for Building a Technical Thesaurus.<sup>1</sup> Drafts of this document was circulated within the DOD community for coordination. A slightly modified version of the guidelines were included in TEST and has been published separately by EJC as Thesaurus Rules and Conventions. The nature of

---

<sup>1</sup> U.S. Office of Naval Research, Washington, D. C., April 1966, 29 pp.  
(AD-633 279)

undertaking, i.e. that the product was to be a model, made it necessary to provide guidelines that would be easy to follow and which would represent current thinking with regard to subject indexing vocabulary requirements. The guidelines were intended (1) to establish a rationale for selecting and displaying the vocabulary of TEST and to promote consistency during its compilation and (2) to aid in utilizing excerpts of TEST in special situations or in compiling separate compatible thesauri. It should be noted that few strict rules are presented. Rather, an effort was made to point up various situations that may be encountered in selecting terms and specifying term relationships and to suggest some factors that should be considered in dealing with these situations.

### Vocabulary

In developing an indexing vocabulary, three separate, but closely related processes are required. They are, (1) the identification of the concepts that are to be represented, (2) the choice of terms to represent those concepts, and (3) the determination of the exact forms of the terms that appear in the formalized vocabulary. As will be seen, these considerations are, to a large degree, interdependent, but it seems convenient to discuss them separately and in roughly the reverse of their order of importance. References are provided to the sections of the conventions that appear in TEST.

### Term Construction

For the most part, term construction in TEST was dictated by usage. Primary emphasis was placed on presenting terms in the manner in which they appear in the literature, subject to a few considerations of machine processing, parallelism, and definition. Considerations of term construction stem from more than a concern for editorial niceties. Some term displays in TEST are dependent upon consistency in the actual forms of the terms.

Although the literature is replete with abbreviations, initialisms, and acronyms, an effort was made, in the interest of communication, to keep the use of constructions of these kinds to a minimum in TEST (Paragraph T-9). A few abbreviations were used when it seemed reasonably certain that their meanings would be understood, but even in these cases the abbreviations were cross referenced from their spelled out forms. Because of their transitory nature, only the most commonly used acronyms and initialisms were included. In the interest of simplicity and consistency, and to facilitate computer processing, punctuation was kept to a minimum (T-8). A maximum term was set at 34 characters to accommodate a three column format.

The application of the foregoing principles was more or less mechanical and presented few problems in compiling TEST. However, some conventions were used that represented departures from the term construction of other thesauri or that presented problems in application.

The prescribed use of direct entry (T-4) raised some objections because, in some vocabularies, indirect entries are used to bring together terms having a generic word in common. It was decided that the cross reference structure and the Permuted Index would display the same kinds of relationships as indirect entries and would perhaps be more useful. Moreover, indirect entries present a great many opportunities for inconsistency in term construction and can lead to confusion. The device of providing parenthetical qualifiers for some ambiguous terms (T-5(b)) results in some constructions that are similar to indirect entries but this proved to be a minor problem.

For reasons that are not entirely clear, the preferred use of plural forms (T-3) was an unexpected source of objections, although the guidelines appear to be consistent with the practice that has prevailed in subject authority lists for some time. There are a number of operational thesauri in which singular forms are preferred. It would be difficult to demonstrate a significant advantage of either approach over the other. The experience at Project LEX was that the preference for plural forms presented no particular problems, despite that apparent complexity of the rule as set forth in the conventions. An exception to this preference for plurals was permitted in the choice of some terms to represent parts of the body. This was done in the belief that it was consistent with common usage in the medical field.

The vocabulary of TEST was limited to terms that can stand alone to represent valid subjects for indexing. Hence, adjectives and terms that are used in the same way as adjectives were excluded (T-2). This is significant in that it represents a further step away from the reliance upon term coordination which characterized early thesaurus development and which is still evident in some operational thesauri. The intent was not necessarily to deny the validity of vocabularies designed expressly for coordinate retrieval, but to insure that the model would be one of the widest possible applicability. For much the same reason, the decision was made to exclude verb forms and to express processes, actions, and the like as gerunds or as some similar noun form.

#### Term Selection

The main factors governing term selection for TEST were usage and definition. Where terms in common usage were in some way ambiguous, the intended meaning was spelled out for a parenthetical qualifying expression, by inserting a preceeding adjective, or if necessary by a scope note. The intent was to make each term completely unambiguous, but efforts to relieve ambiguity sometimes conflicted with efforts to adhere to usage and to maintain consistency in term construction. The policy followed in TEST was to compromise on usage and construction when necessary in order to relieve ambiguity. For example, subtle differences in term construction were found to be insufficient as a means of distinguishing between some concepts. It should be noted that the principle set forth in T-5(c) was not followed.

Apart from questions of ambiguity, a considerable number of commonly used terms are, for various reasons, not to be interpreted literally. Terms of this

kind were not altered for inclusion in TEST, but effort was made to take their definitions into account for cross referencing purposes.

It may be appropriate here to interject two additional points that the experience at Project LEX brought out.

In compiling the thesaurus, reference materials must be relied upon extensively to determine the proper definitions of terms and to decide upon preferred usage. The many subject experts that participated in compiling TEST provided invaluable insight in selecting important concepts and suggesting terminology, but they too were often enlightened by referring to glossaries, encyclopedias, and other authorities. Of course, reference works can disagree, but almost without exception the usage in TEST can be traced to a published authority.

Thesaurus builders are presented many temptations both to coin terms for special situations, such as to fill levels in hierarchies or to make hierarchies appear symmetrical, and to revise some commonly used terms that appear to be misnomers. These temptations must be resisted. Synthesized terms, no matter how appropriate they may seem, have a way of becoming meaningless in indexing and retrieval. There are coined terms in TEST, particularly in the fields of chemistry and metallurgy, but these were carefully chosen and have been explained by scope notes. As a rule, no term should be used that cannot be substantiated by a separate authority, but if a coined term is absolutely necessary, its meaning must be made perfectly clear.

#### Concept Identification

The most vexing aspect of compiling TEST was the proper choice of the concepts that were to be represented. Paragraphs T-7 and T-11 of the conventions deal with this problem, although the presentation may be somewhat misleading. Both paragraphs refer to choosing among terms, but the heart of the matter is the identification of appropriate concepts for indexing and retrieval. The two paragraphs are actually discussions of decisions that must be made rather often in selecting appropriate concepts.

The discussion of what are called "quasi-synonyms" (T-7) deals with an important aspect of concept identification, although the wording may be unclear. What is meant is that once an indexing concept has been identified, subjective judgments may indicate that one or more other concepts, although different in some way, are so similar in meaning that no distinction need be made for indexing and retrieval purposes. Moreover, some concepts cannot be distinguished, although various terms may be used in the literature when different points of view are involved. The decisive factor, then, is the determination of the valuable and useful indexing concepts. If the guideline is construed in this way rather than as a term selection problem, indexing effectiveness and consistency will be served without hampering healthy vocabulary growth.

The intent of paragraph T-7 is to deal with concepts that are related in ways other than generically. The intent of paragraph T-11 should be to deal with the proper specificity of concepts represented in TEST. What is otherwise a rather concise presentation of the problem of concept specificity is obscured by the consistent use of "multiword term" where "specific term" would explain the more general problem. The confusion comes about from the preoccupation with coordinate indexing applications mentioned earlier. So much power was attributed to coordination in indexing and retrieval that some authors have apparently viewed terms

comprising two or more words as "precoordinated" and therefore odious. The wording of T-11 reflects an overreaction to this misconception. The number of words that comprise a specific term must not be allowed to cloud the issue. A determination must be made of the concept that is to be represented and how, in the context of the vocabulary as a whole, that concept can best be represented.

Certainly, many very specific concepts are best represented by very specific terms that happen to contain several words; others may be represented by single words. The same decisions must be made in both cases. Once the indexing and retrieval value of the concept has been established and the appropriate term chosen, then the relationships, if any, to concepts and terms in the thesaurus vocabulary can be determined and the best means of representing the concept can be decided upon. At this point, the considerations set forth in T-11 should be made, provided they are rephrased as follows: a specific term must be established when no more general term is available; a specific term should be established when the specific concept is very important in the operational situation; consideration should be given to the use of two (or more) more general terms to represent the specific concept when each general term represents a concept generically related to the specific concept; and, the specific term should be established when doubt remains after the foregoing have been considered.

The conventions make no mention of the possibility that a specific concept can be indexed by a single term representing a slightly more general concept, though this is prescribed by entries in TEST. In fact, there appear to be far more instances in which simple generalizations were made than there are of the other kinds of term selection. In a given application, some of the decisions reflected in TEST will no doubt be rejected, but the important judgment regarding concept identification and term selection must be based on the criteria set forth in the conventions.

#### TEST Format

The TEST format is quite similar to that of the EJC Thesaurus, except that three indexes, or vocabulary displays, have been added. The format is explained at some length in the introductory material, so need not be described here. Rather, mention will be made of a few considerations relating to the development of the format and its use in organizing the vocabulary.

#### Thesaurus of Terms

The Thesaurus of Terms, or alphabetical section, follows the EJC format almost exactly. The format appears to be useful and generally well accepted and is one of the least complicated of existing thesauri. This section is intended to be the principal vocabulary tool; the indexes should be considered adjuncts to it. The Thesaurus of Terms lists the entire vocabulary in alphabetical order with cross references to show relationships among terms. The relationships are established on the basis of the definitions of the terms. As was mentioned earlier, it is necessary first to identify an indexable concept, then to determine what term best represents that concept, and finally to specify what relationships the term has to other terms in the thesaurus. The two defineable relationships specified in TEST are synonymy and quasi-synonymy, shown by USE and used for (USE-UF) references, and class membership, shown by Broader Term-Narrower Term (BT-NT) references. The Related Term (RT) references are not defineable but are developed subjectively.

The USE-UF reference (C-2, C-3), as might be inferred, is employed to show a preference between synonyms, to show where quasi-synonymy has been found, to prescribe a combination of terms to represent a concept, or to show where a generalization has been made.

The BT-NT reference (C-6) is employed in every case where an invariable class relationship between terms exists, that is, where one term represents a class and a second term represents a member of that class. The most important consideration is ascertaining that the relationship is one of true class membership and not a part-whole or class of use relationship. Some exceptions, notably for anatomical terms, were made to this rule in TEST, but it would seem better to have applied it consistently.

The selection of RT references (C-8) presented some difficulties. The need for cross referencing among terms that are related in certain undefineable ways is generally acknowledged, but there seems to be no way to maintain a consistent approach. Furthermore, the viewpoints of thesaurus users cannot be anticipated, so that some RT's will appear superfluous to some users and, perhaps, useful RT's will be omitted. Probably TEST errs on the side of superfluous RT's.

Each set of cross references was always made reciprocal (C-9) and all levels of a hierarchy are shown at each entry. This means that, for example, a general term may have a large number of NT's that represent several levels of specificity. In such a case, the Hierarchical Index should be consulted to obtain a more intelligible display. This redundancy among NT's helped avert errors in hierarchies during the compilation of TEST, and will probably be an aid in excerpting portions of the vocabulary. Some reciprocal RT's are superfluous, particularly those from specific terms to relatively general ones. These, too, were deliberately included to aid in editing the vocabulary.

#### Permuted Index

The Permuted Index, essentially a computer sort or KWIC index of the words in the vocabulary, proved to be extremely useful in the final indexing phase of Project LEX and is expected to be even more valuable as an aid to thesaurus users. Since each word in each term is an entry point, all terms having words in common file together and provide a collection point for terms that are separated because of the use of direct entries.

#### Hierarchical Index

The Hierarchical Index displays the BT-NT relationships for all terms. It will probably be most useful in retrieval, particularly in mechanized systems that have hierarchical search capabilities. In addition, it provides an orderly display of the more complex hierarchies.

### Subject Category Index

The Subject Category Index will be of use in indexing and retrieval when it is necessary to determine generally the scope or depth of vocabulary development in some subject area. The most common application is expected to be in segmenting TEST, such as might be done in constructing specialized thesauri. It should be pointed out that, although effort was made to conform to the COSATI Subject Category List, several departures were required. The resulting displays are believed to be reasonably coherent and of useful content, but the real utility of this display has not been determined.

### Alphabetization

The matter of alphabetization (A-1) consumed an inordinate amount of time in developing the conventions. Suffice it to say that the relative merits of word-by-word vs. character-by-character arrangement were discussed at length and from many points of view before the latter was adopted.

### Conclusions

The statement of conventions was extremely useful in the compilation of TEST. There were some instances in which specific provisions of the conventions were ignored or revised in practice, but for the most part, the guidance was found to be sound and was followed. The use of these conventions, or an adaptation of them, is recommended as a starting point in any thesaurus compilation effort similar in nature to the development of TEST.



## SATELLITE THESAURUS CONSTRUCTION

William Hammond

Automated Systems Corporation

### Purpose

It is assumed that the publication of scientific and technical information thesauri for the large government information facilities has had a profound effect on the entire technical information community. If this were not the case, it is doubtful that two years after publication of the Department of Defense Thesaurus of Engineering and Scientific Terms (TEST) this panel discussion would be on the agenda. Speculation at this point in time on the pros and cons of the effectiveness of a thesaurus controlled indexing vocabulary for information retrieval is irrelevant. To paraphrase the diplomat's prayer, let us at least hope that it does no positive harm -- and that the overall benefits to be derived by the adoption of compatible terminology will far outweigh any adverse aspects of a rigidly controlled vocabulary. Anyway, the decision to produce and employ thesauri in the management of scientific and technical information was made by those in authority some time ago. This paper is concerned with how to make the most of this decision.

### Background

Those of you who are concerned with the operation of a technical information system must determine the extent -- if any -- to which all or part of a given thesaurus will be incorporated into your system. If any significant subset is to be incorporated, a machine capability to handle the "bookkeeping" will be a great asset. In most instances, existing thesauri can be obtained on magnetic tape. Government agencies may obtain copies of the DOD TEST on tape from the Defense Documentation Center. Non-government agencies may purchase the TEST tapes from the Engineers Joint Council. Other Thesauri may be obtained in most instances from the originator.

It was intended that this panel center its discussion around the DOD Thesaurus and its use. It is relevant here to review the Project LEX effort to put its "creation" in proper perspective. The Department of Defense spent more than one-half of a million dollars from appropriated funds for Project LEX. Additionally, 328 volunteer panelists contributed their time and bore their own travel and incidental expenses. Among the 328 volunteers were 46 Ph.D.'s representing almost every scientific discipline. The vocabularies from 140 or so different operational information systems were assembled into a common-format, composite data store on

magnetic tape. Computer manipulation of the data store produced groupings of the terms by subject and source as well as permuted and hierarchical arrays. These end products were scrutinized by the panelists to develop candidate terminology. Many of the major thesauri constructed since Project LEX have in varying degrees made use of this prodigious LEX effort.

The COSATI 1 September 1967 publication, Guide Lines for the Development of Information Retrieval Thesauri, is substantially the guide lines published earlier by the Engineers Joint Council and later modified somewhat to govern the LEX effort. The COSATI version of the guidelines is more permissive to some extent, particularly in the alphabetical sequencing of terms in the published thesaurus. This has an important bearing on computerized uses of the thesaurus corpus.

### Thesaurus File Structure

The format of the TEST magnetic tape file is contained in Attachment 1 to this paper. A "dump" of a portion of the tape file is reproduced on Attachment 2. Several other thesauri use a compatible variation of the TEST magnetic tape layout. These include NASA Thesaurus, Urban Thesaurus, Fort Detrick Thesaurus, S.C. Johnson (Johnson's Wax) Thesaurus, Linguistics Thesaurus and the Department of State Thesaurus, and others.

From the dump of the TEST file it is obvious that more detailed information is needed for computer manipulation. This supplemental information can be obtained from one of the two sources given earlier. The file contains redundancy in thesaurus line codes and term sequencing codes in addition to the repetition of the main term entry for each of the cross references. This format was a carry over from the file format adopted to accommodate the LEX data store references earlier. It has since proven to be quite efficient for thesaurus updating and for manipulating the thesaurus corpus by computer to produce various term displays. The file organization is also convenient for construction of "Satellite" thesauri which might be looked upon as an update in so far as the computer processing is concerned.

### Thesaurus Model

What a thesaurus is and is not has been covered quite adequately by other panel members. To the legal mind the terms structured in the thesaurus provide only circumstantial evidence of a document's subject content. The thesaurus, however, goes beyond the dictionary's single words to define multi-word terms or phrases that greatly increase the potential specificity for describing subject content. What has evolved from this quest for more and more precise indexing terminology provides, for the first time, an indexing vocabulary structure or "model" that is quite susceptible to computer correlations that are associated with the meanings of terms -- meanings within the constraints of the context of the thesaurus.

There is a unique definition for each term in a thesaurus that is constructed to the COSATI guide lines. This definition is embedded in an explicit term display that includes one or more subject categories to which the term is relevant; a limiting scope note, if needed; a set of cross references to all other terms in the thesaurus that are synonyms, broader, narrower, or related conceptually.

A glance at the example from the NASA Thesaurus in Figure 1 will show that "Radiation Spectra" is one of the (14,940) postable NASA terms; its definition overlaps four subject categories (numeric codes only shown). Among the other 14,939 postable terms in the NASA vocabulary, only one term is broader than Radiation Spectra; however, twenty-six postable terms are narrower and five are conceptually related. Added definition can be gleaned from the different term displays that are published as indexes to the thesaurus.

Satellite thesaurus construction procedures can best be described within the framework of the case history of the Fort Detrick Thesaurus which combined a list of 6,000 "authorized descriptors" from a seasoned, computer based information system with the TEST corpus.

The Fort Detrick terms were keypunched and used in the computer to "pull" matching TEST terms together with all their cross references from the TEST magnetic tape file. The embryo thesaurus corpus thus produced retained two types of TEST entries: Fort Detrick/TEST matching terms with their entire cross reference set from TEST and other TEST terms that were cross referenced to authorized Detrick terms. In the latter entries only the cross references to the Detrick terms were retained.

In the first pull from the TEST file, about 4,000 Detrick terms matched those in TEST; however, by pulling the non-authorized or "first generation" cross referenced entries, about one-half of the TEST corpus was pulled! In a second review of its authorized terms, Detrick was able to substitute TEST terminology for all but 1,100 of its original 6,000 terms. Many of this 1,100 residue were project names and nomenclature of the type intentionally omitted from TEST.

Fort Detrick decided to maintain the entire composite TEST/Detrack magnetic tape file with tags to permit selective compilations from a single file to display the full composite corpus; the Fort Detrick corpus only, with cross references only among its authorized terms; and the Fort Detrick corpus plus the TEST terms cross referenced to the Detrick terms.

It was necessary for Detrick to establish cross references among its own terms not listed in TEST and between these non-TEST terms and terms in TEST. Cross references were established only for synonyms (USE), immediate broader term(s) if any, and any related terms (RT). A computer pass generated reciprocal cross references and filled out the intermediate generic (BT-NT) structure from the immediate BT "thread" established intellectually. Essentially, Fort Detrick retained the TEST format. Modifications were made to the numeric term identification codes and tags were added to identify Detrick authorized terms and their cross references from TEST.

RADIATION SPECTRA  
1411 2402 2406 2902 2903  
BT #SPECTRA  
NT ABSORPTION SPECTRA  
BALMER SERIES  
D LINES  
ELECTROMAGNETIC SPECTRA  
ELECTRONIC SPECTRA  
EMISSION SPECTRA  
FRAUNHOFER LINES  
H ALPHA LINE  
H BETA LINE  
H GAMMA LINE  
H LINES  
HERZBERG BANDS  
INFRARED SPECTRA  
K LINES  
LINE SPECTRA  
LYMAN SPECTRA  
MICROWAVE SPECTRA  
PASCHEN SERIES  
RADIO SPECTRA  
RAMAN SPECTRA  
RYDBERG SERIES  
SOLAR SPECTRA  
STELLAR SPECTRA  
TELLURIC LINES  
ULTRAVIOLET SPECTRA  
VIBRATIONAL SPECTRA  
RT ASTRONOMICAL SPECTROSCOPY  
ENERGY SPECTRA  
MASS SPECTRA  
NOISE SPECTRA  
PLASMA SPECTRA

Figure 1

There are many advantages to retaining the TEST file format. The most important is that subsequent updates of TEST can be accommodated in a computer update that can also flag conflicts in term relationships between the TEST cross references and cross references in the satellite thesaurus. The full BT-NT display carried in TEST also permits a computer diagnosis of the hierarchical structure and to generate the hierarchical displays to aid the indexer or retriever as well as detect deficiencies in the thesaurus structure.

In establishing the BT threads (immediate broader term) for the satellite thesaurus, it is very helpful to make a final review with two tests for each term:

- o Is this term one of these BT's?
- o Should this term and its NT cross references be listed as NT's under the BT entry?

Although these two questions appear to be over simplifications, they have proven to be quite useful. Reviewing from the BT cross reference is recommended simply because it represents a less complex array in TEST. There is seldom more than one BT for each hierarchical level in TEST; however, it will often list several dozen NT's of the same level under a given term.

In the last analysis it must be remembered that under the "thesaurus concept" a word means whatever one wants it to mean, nothing more, nothing less. This holds true so long as it is possible to define the given word by the display (or omission from the display) of its relationships with the other words listed in the thesaurus.

## DESCRIPTION OF LEX MAGNETIC TAPE LAYOUT

A	B	C	D	E	F	G	H	I	J	K	L	M

109 CHARACTER RECORD -- 13 FIELDS -- IBM MODE \* 1)

FIELD CONTENT

<u>FIELD</u>	<u>POSITION(S)</u>	<u>CONTENT</u>
A	1-7	Reserved * 2)
B	8	Term relationship code; see code key below
C	9-14	Reserved
D	15	Line sequence code for scope note; type of pseudo scope note; see code key below.
E	16-19	Reserved
F	20-21	Reserved
G	22-23	Reserved
H	24	Reserved for term tag * 3)
I	25-60	36-character term entry, scope note line, or subject category codes; this field is also referred to as the <u>sub-term field</u> ; see key below.
J	61-66	Reserved; used for carrying numeric surrogate of sub-term in Field I (25-60) or for extending capacity of sub-term field from 36 to 42 characters.
K	67-102	36-character term entry; this field also referred to as <u>main-term field</u> .
L	103-108	Reserved; used for carrying numeric surrogate of main term in Field K (67-102) or for extending capacity of main-term field from 36 to 42 characters.
M	09	Reserved; used for record indicator when required for computer configurations other than IBM 360.

"Reserved": Unless otherwise specified, the content of reserved fields varies with the application.

LEX RECORD IDENTIFICATION CODE KEY

- . Main-term record (MT) is identified by a 1 in position 8.
- . Scope Note (SN) is identified by a 2 in position 8 and a numeric line sequence code (1 through 9) in position 15.
- . COSATI Subject Category Code is identified by a 2 in position 8 and a C in position 15. Each subject category is represented by a set of 4 numeric digits, beginning in position 25, with a blank between each 4-digit set. Thus, seven category codes may be carried in a single record. If more than seven codes are required, an additional record is formed.
- . A "USE" cross-reference is identified by a 3 in position 8.
- . A "Used for" (UF) cross-reference (reciprocal of USE) is identified by a 4 in position 8.
- . A "Broader term" (BT) cross-reference is identified by a 5 in position 8.
- . A "Narrower term" (NT) cross-reference is identified by a 6 in position 8.
- . A "Related term" (RT) cross-reference is identified by a 7 in position 8.

\* 1 800 bpi, 9 Track, 300 records per block

\* 2 7-digit numeric line sequence code


\* 3 System/360, 8-bit code shown below indicates:

01001011 = narrower terms listed in thesaurus

01001110 = refer to main term entry

00000011	Abaca fibers	000100Abaca fibers	000100
00000023	Manila hemp	371050Abacca fibers	000100
00000031	Abandonment	000130Abandonment	000130
00000042	1407	000130Abandonment	000130
00000057	Depletion	170530Abandonment	000130
00000067	Depreciation	170860Abandonment	000130
00000077	Escape (abandonment)	220120Abandonment	000130
00000087	Life (durability);	350530Abandonment	000130
00000097	Maintenance	368770Abandonment	000130
00000107	Oil wells	430210Abandonment	000130
00000111	Abatement	000160Abatement	000160
00000122	1407	000160Abatement	000160
00000137	Attenuation	052030Abatement	000160
00000147	Damping	162430Abatement	000160
00000157	Dispersing	183730Abatement	000160
00000167	Disposal	184090Abatement	000160
00000177	Dissipation	184300Abatement	000160
00000187	Exhausting	224410Abatement	000160
00000197	Pollution	477670Abatement	000160
00000207	Purification	502360Abatement	000160
00000217	Smoke abatement	574270Abatement	000160
00000227	Stopping	602260Abatement	000160
00000231	Abbreviations	000190Abbreviations	000190
00000242	0502	000190Abbreviations	000190
00000254	Acronyms	006940Abbreviations	000190
00000267	Mnemonics	396730Abbreviations	000190
00000277	Symbols	620230Abbreviations	000190
00000281	Abdomen	000220Abdomen	000220
00000292	0616	000220Abdomen	000220
00000307	Digestive system	178750Abdomen	000220
00000317	Peritoneum	453160Abdomen	000220
00000321	Abdominal dropsy	000250Abdominal dropsy	000250
00000333	Ascites	047230Abdominal dropsy	000250
00000341	Abducent nerve	000280Abducent nerve	000280
00000352	0616	000280Abducent nerve	000280
00000365	Cranial nerves	151660Abducent nerve	000280
00000375	Peripheral nervous system	452860Abducent nerve	000280



MEMO ROUTING SLIP		NEVER USE FOR APPROVALS, DISAPPROVALS, CONCURRENCES, OR SIMILAR ACTIONS		ACTION XX	
1 TO Defense Documentation Center ATTN: Mr. Myer B. Kahn, DDC-TC Bldg. 5, Cameron Station, Alexandria 22314		INITIALS	CIRCULATE		
		DATE	COORDINATION		
2			FILE		
			INFORMATION		
3			NOTE AND RETURN		
			PER CONVERSATION		
4			SEE ME		
			SIGNATURE		
<b>REMARKS</b>  Attached two copies of "The Thesaurus in Action" are furnished for accessioning. Demand exceeded our supply at the Convention.  Pages 21 through 25 have been removed. They were included by administrative error in the assembly of the papers, and did not pertain to the subject.  <div style="text-align: center;">  </div> <p style="text-align: center;"> <b>PARMELY C. DANIELS</b>  <b>CRDISO- DATA MANAGEMENT DIVISION</b>              DEPARTMENT OF THE ARMY              Chief, Chief of Research and Development           </p>					
FROM		DATE		4 OCT 1969	
		PHONE			

**DD FORM 95**  
1 OCT 60

REPLACES PREVIOUS EDITION

★ GPO: 1968-O-312-129

## THE USE OF TEST IN THE PREPARATION OF THE RESEARCH AND DEVELOPMENT CAPABILITY INDEX

James G. Peirce

Frankford Arsenal  
Philadelphia, Pa. 19137

### Introduction

The Army information program for potential defense contractors called qualitative requirements information (QRI) has been instrumental in designing the Research and Development Capability Index, a hierarchically structured form to be used in the development of research and development source lists. This form, DD Form 1630, is derived from the COSATI Subject Category List. It is mainly intended for use in DoD and NASA procurement activities. However, the Army has planned it also for use in describing the capabilities of civilian organizations qualified to receive research and development advanced planning data and information.

The form was developed in generally the same time frame, but slightly ahead of the DoD project LEX which produced TEST. It was coordinated with TEST activities. Its language was one of the inputs to TEST. The final refinement and adoption of the DD Form 1630 by the Armed Services Procurement Regulations (ASPR) in 1967 met an industry request to DoD from the National Security Industrial Association (NSIA).

The development of an Army thesaurus for the DD Form 1630 language was a logical related development. Factors involved in this activity are described. The plan for the thesaurus and its present contents show their close affinity to TEST and the COSATI list. This thesaurus will become an automatically updated open-ended language file for the QRI Registered Organization Data Bank (RODATA), the computerized retrieval system being designed for control of civilian responses to QRI and other unsolicited R&D proposals. It will be published by early April 1970 as an appendix of DA PAM 70-20-1.

This is the third of a series of papers on system developments in the Army program for providing information services for potential defense contractors. The expanded program definition embraces the entire spectrum of guidance information that the Army can release to qualified civilian sources in advance of specific procurement requirements, for planning purposes. In 1967 Peirce and Shannon (1) described the cartridge type microfilm system selected for uniform storage of registrants' qualification data. In 1968 Peirce and Segal (2) described the Army's planned implementation of source data collection utilizing the DD form 1630 capability index. Today I am going to describe the Army effort that went toward the establishment of organizations offering resources and research and development services to DoD.

### COSATI, LEX AND QRI

The COSATI Category List (AD 624 000) (3) has been accepted by all DoD agencies as a basic tool for classification schemes and the language for technical information retrieval. Older Air Force and Army forms had been established for research and development source list and bidder's list operations. These were based on the old ASTIA Information Guide. The same pressures that led to the production of the COSATI list led to the development of a revised classification language for recording research and development capabilities, which activity was initiated early in 1965. By this time the December 1964 first edition of the COSATI Subject Category List had been published, and the planned use of this list for Defense Documentation Center (DDC) classification and retrieval of technical documents was known. It was not difficult to decide that the COSATI list would become the new base on which to build the Army classification language for qualification of civilian organizations as potential contractors. An Army (AMC) language classification committee (4) was organized which met for eighteen months developing a new COSATI - compatible classification scheme. Representatives of major installations in the Army Materiel Command were assigned responsibilities for the twenty-two COSATI fields as indicated by Figure 1. About every two months, during the last half of 1965 and the whole of 1966, the entire committee reassembled to review progress and results. It was usually at this time that installation representatives with contributions to other than their assigned fields would contribute individual terms to the monitoring agency, or to a select coordinating team (including contractor personnel) that circulated during work sessions.

Consideration was simultaneously given to the development of a revised form. The format the Army designed is now the current DoD format. It is also in use in several places for recording data of local interests where use of the form is not applicable. Planned for production later are revised capability index sheets designed for use with optical readers. Army developed terms were submitted to Project LEX and have become part of the TEST vocabulary. Also, the conventions established for LEX were generally adopted as written for regulating the various inputs to the R&D Capability Index. Although all of the terms finally adopted by the Army did not match LEX terms exactly, all were submitted to LEX scrutiny and a relationship was established to a LEX term, where possible. The Army and LEX approaches were closely coordinated.

By early 1966, at the time of a committee meeting at Redstone Arsenal, the total concept of the Army vocabulary was clearly evident. The submissions from the various AMC agencies were extremely varied. They included repeats of the terms that appear usable from the DD Form 558-2, revised terms to replace cumbersome 558-2 terms, and new terms which had come into use since the organization of the original ASTIA based language. The relationship of DD Form 558-2 divisions to the DD Form 1630 (COSATI) fields and groups is shown in Figure 2. DDC supplied the AMC Committee with a duplicate set of the punched cards which had produced the alphabetic index published in the October 1965 edition of the COSATI Lists - DDC Expanded (AD 624 000).<sup>\*</sup> The terms submitted by each installation were converted to machine readable form by the DURAMATIC Army Chemical Typewriter Mach II

# COSATI Field Coordination Assignments

## Missile Command

- 3 - Astronomy & Astrophysics
- 10 - Energy Conversion
- 16 - Missile Technology
- 18 - Nuclear Science
- \*21 - Propulsion & Fuels
- 22 - Space Technology

## Electronics Command

- 4 - Atmospheric Sciences
- 9 - Electronics & Electrical Engineering
- \*15 - Military Sciences
- 17 - Navigation, Communications, Detection & Countermeasures
- \*18 - Nuclear Sciences & Countermeasures
- 20 - Physics
- \*22 - Space Technology

## Mobility Equipment R&D Center

- 2 - Agriculture
- 8 - Earth Sciences & Oceanography
- 13 - Mechanical, Industrial, Civil, & Marine Engineering
- 17 - Navigation, Communications, Detection & Countermeasures
- \*19 - Ordnance
- 21 - Propulsion & Fuels

## Frankford Arsenal

- \*9 - Electronics & Electrical Engineering
- \*10 - Energy Conversion
- \*17 - Navigation, Communications, Detection & Countermeasures
- \*20 - Physics

## Weapons Command

- 12 - Mathematical Sciences
- \*14 - Methods and Equipment
- 19 - Ordnance

## Edgewood Arsenal

- 5 - Behavioral & Social Sciences
- 6 - Biological & Medical Sciences
- 7 - Chemistry

## Munitions Command

- 11 - Materials
- 21 - Propulsion & Fuels
- 15 - Military Sciences

## Aviation Command

- 1 - Aeronautics
- \*8 - Earth Sciences & Oceanography
- 17 - Navigation, Communications, Detection & Countermeasures

## Tank Automotive Command

- \*13 - Mechanical, Industrial, Civil, & Marine Engineering
- 14 - Methods & Equipment

\*Secondary Responsibility

FIGURE 1. Field Assignments Made to AMC Installations

**DD Form 558-2 Divisions**

1. Aircraft & Flight Equipment
2. Astronomy, Geophysics, & Geography
3. Chemical Warfare Equip & Materials
4. Chemistry
5. Communications
6. Detection
7. Electrical Equipment
8. Electronics & Electrical Equip
9. Fluid Mechanics
10. Fuels & Combustion
11. Ground Transportation Equipment
12. Guided Missiles
13. Installations & Construction
14. Materials (Nonmetallic)
15. Mathematics
16. Medical Sciences
17. Metallurgy
18. Military Sciences and Operations
19. Navigation
20. Nuclear Physics and Nuclear Chemistry
21. Nuclear Propulsion
22. Ordnance
23. Personnel & Training
24. Photography & Other Repro Processes
25. Physics
26. Production & Management
27. Propulsion Systems
28. Psychology & Human Engin
29. Quartermaster Equipment & Supplies
30. Research & Research Equipment
31. Ships & Marine Equipment
32. Miscellaneous Arts & Sciences
33. Transportation
34. Bio-Astronautics
35. Spacecraft & Space Equipment
36. Range Operations & Studies

**DD Form 1630 Fields**

1. Aeronautics
3. Astronomy & Astrophysics
8. Earth Sciences & Oceanography
15. Military Sciences
7. Chemistry
17. Navigation, Communications, Detection & Countermeasures
17. Navigation, Communications, Detection & Countermeasures
9. Electronics & Elec Engin
9. Electronics & Elec Engin
20. Physics
21. Propulsion & Fuels
13. Mechanical, Industrial Civil and Marine Engin
16. Missile Technology
13. Mechanical, Industrial, Civil and Marine Engin
11. Materials
12. Mathematical Sciences
6. Biological and Medical Sciences
11. Materials
15. Military Sciences
17. Navigation, Communications, Detection & Countermeasures
7. Chemistry
18. Nuclear Science & Technology
20. Physics
21. Propulsion & Fuels
19. Ordnance
5. Behavioral & Social Sciences
14. Methods & Equipment
20. Physics
5. Behavioral & Social Sciences
13. Mechanical, Industrial, Civil and Marine Engin
21. Propulsion & Fuels
5. Behavioral & Social Sciences
15. Military Sciences
14. Methods & Equipment
13. Mechanical, Industrial, Civil and Marine Engin
5. Behavioral & Social Sciences
15. Military Sciences
6. Biological & Medical Sciences
22. Space Technology
15. Military Sciences
22. Space Technology
14. Methods & Equipment
16. Missile Technology

Figure 2

Correlation Between DD Form 558-2 and COSATI Fields

and then merged into a single printout listing with the DDC list. A series of meetings at Natick Laboratories, the Weapons Command, and Ft. Belvoir completed selection of a total Army list arranged in COSATI group structure with two additional tiers of data tentatively called "Sections" and "Units." A coding system consistent with the number codes assigned in the DDC expanded COSATI list was assigned to the new terms at the section and unit levels. It was found necessary to add relatively few terms at group level (see Figure 3). These additions were coordinated with similar recommended changes being proposed by Project LEX. By the end of November 1966 a new punched card listing had been prepared and a final committee exercise was run to complete Project LEX coordination.

#### Participation in Project LEX

In addition to the coordination of Army language with Project LEX, arrangements were also made in 1966 for active participation in LEX activities. LEX scheduled a series of basic field work sessions which were directly relatable to the COSATI field review assignments of our committee. The AMC committee members were asked to schedule knowledgeable persons from their installations for LEX participation. About ten language scientists or engineers expert in the language of their chosen fields were obtained from this call (5). For instance, three representatives of the Munitions Command attended the opening session of LEX on Materials at the New York City headquarters of the Engineers Joint Council (EJC) in May 1966. Others are mentioned in TEST's list of participants.

In the early winter of 1966-67 the initial LEX permuted index printout was checked out against the final selection of Army terms. This was done at a Washington, D.C. meeting in December 1966. In advance of the meeting the Army terms were listed alphabetically to facilitate faster checking against the LEX permuted index. The Army list was divided into five parts so that teams of two could review it. More than half of the Army terms were found to be already in agreement with LEX selections. Another 25%, approximately, which clearly agreed with LEX in meaning, were edited into the LEX preferred format. For about 10% of the terms, not quite half of the remainder, the Army selection of phraseology, although apparently identical in meaning to LEX terms, was not to be altered to the equivalent LEX phrase in the opinion of those best qualified to recognize how the term was used. Such Army submissions were allowed to stand. The last 15% of terms were those which could not be matched at the meeting. As committee chairman, later in December I spent four days at the Project LEX offices reviewing this last group of terms. Equivalents were found for almost 80% of this group. However, in consideration of customary Army usage, only 30% were modifiable to LEX terminology. After this review process the terms were resorted into the COSATI based hierarchical structure, and final modifications were made in coding assignments caused by alphabetical changes in the review operations. In summary the following situation existed after the LEX-QDRI comparative review:

Terms in original agreement	52%
Changed, LEX terminology adopted	
Committee review	25%
Chairman's review	5%
Has LEX equivalent, but not changed	10%
No LEX equivalent	8%

Table 1

Results of LEX Comparison and Edition, December 1966

### The ASPR Subcommittee

In October 1966 an ASPR (Armed Services Procurement Regulation) Subcommittee was charged with the establishment of a uniform DoD survey form for industrial research and development capabilities. This subcommittee was also to recommend changes to appropriate ASPR's, and to coordinate its activities with NASA (at Goddard Space Flight Center). Through the National Security Industrial Association (NSIA) industry had already gone on record as desiring, almost demanding, such a uniform approach. This form was basically to be designed as a bidders mailing list classification tool; however, the Army took the position that the same item would be used for both bidders mailing lists and the research and development information program. The eighteen months prior work by the Army QDRI language committee now became the Army input to the joint form.

In addition to the four-tiered Army list, already established, the subcommittee had available to it lists of terms prepared by the Air Force, Navy, NASA and NSIA. The Air Force, Navy and NASA terms were merged into a single list by AFSC. The Army then took over the merge of this list with the Army and NSIA contributions. Originally all terms were merged as separate punched card decks, individually numbered. They were then sorted alphabetically in the tiers to which their submitters had assigned them. Of course there were many duplicate terms. An installation or activity code was given originally to each term, and as duplicate nomenclature cards were deleted this code was added to the resultant card. A sample of the list of codes is provided as Table 2. The easiest duplicates to eliminate were the common terms selected from the COSATI list for the first two tiers. These very easily sorted out next to each other due to identical wording and coding. A small amount of visual search was required for new terms added at the Group level. The great mass of terms were added at the two additional levels. Here, after elimination of duplicates, was another visual task which was performed by the entire subcommittee. Three actions were taken. Each subgrouping was looked at as a whole to consider whether it completely or sufficiently defined the higher tiered term it was qualifying. Then terms with identical meanings were reviewed and one was selected for retention. Some compromises were made and often the final term represented an edited or rewritten phrase which no one agency or installation could claim as a specific submission.

The final word and term selections were returned to the Army for recoding. This was accomplished in about one week's total elapsed time. The subcommittee had accepted the Army form design. This was combined by a QRI contractor operation with the merged list and about 150 copies were printed for submission to the ASPR committee and subsequent planning use by the subcommittee members. The Army finally published the DD Form 1630 with a 1 November 1967 date. Final ASPR approval is in Armed Services Procurement Supplement - ASPS No. 4 dated 1 April 1968.



Installation CodeRODATAInstallationMailing Symbol

01	QRI Committee for Common Scientific Language-COSATI	SMUFA-A2100
02	National Security IND Assoc (RADAC)	RADAC
03	Air/Force - Andrews AFB - AFSC	SCKAE
04	Project LEX & DDC	DDC-DTI
05	Navy-Air Materiel Command	NAIR
10	Hq, US Army Materiel Command	AMCRD
16	US Army Research & Development Center, Aberdeen	AMXRD
20	Natick Laboratories	AMXRE
25	USA Materials & Mechanics Research Agency	AMXMR
30	Harry Diamond Laboratories	AMXDO
40	USA Electronics Command	AMSEL
50	USA Missile Command	AMSMI
60	USA Mobility Command	AMSMO
61	USA Tank-Automotive Command	AMSTA (SMOTA)
62	USA Aviation Materiel Command	AMSAV
63	USA Aviation Materiel Labs	SAVFE
64	USA Mobility Equipment Command	AMSME
65	USA Mobility Equipment R&D Center	SMEFB
70	USA Munitions Command	AMSMU
80	USA Weapons Command	AMSWE
84	San Francisco Procurement Agency	AMXNP
85	Los Angeles Procurement Agency	AMXSP
86	New York Procurement Agency	AMXNY
87	Chicago Procurement Agency	AMXCH
88	Cincinnati Procurement Agency	AMXCN
91	USA Test and Evaluation Command	AMSTE

Table 2

RODATA Codes &amp; Abbreviations

### The Plan for an Army Dictionary

As the body of a new DoD form started to take shape it was generally agreed by the members of the Army committee that some dictionary type efforts were needed in addition to the establishment of a hierarchically structured form. At the January 1966 Washington meeting of the QDRI language committee these requirements were given definition in a plan for a dictionary in four parts, as follows:

1. Scope notes of the third and fourth tier terms;
2. A listing of terms by installation interests, by installation;
3. A similar listing, but alphabetically by terms; and
4. A cross reference of the new form.

Work on this task was assigned to a contract operation in the fall of 1967. Requirements for an alphabetical listing and cross reference of DD Form 558-2 terms were combined into a single permuted index task.

The scope notes were completed finally in original draft form in January 1969. It appeared that the cross referencing between like terms and terms with more than one meaning was quite inadequate. The contractor was asked to rectify this situation and also to simplify quite a few scope note definitions. At the same time agreement on a final format for the permuted index was established (see Figure 4). The contractor was told to use Webster's International, and TEST (the DoD Thesaurus of Scientific and Technical Terms), for both meanings and style guidance.

Two major reviews have been conducted on this Army Dictionary: The AVCOM meeting on the registration process in April 1968, and a total Army review at the Army Research Office, Washington, D.C. in February 1969.

### Future Plans

Although the main immediate purpose of the QRI thesaurus or dictionary is to provide definitions for the terms used for registration classification and a visual cross-reference look-up instrument for converting an older classification language term to a more current one, the fact that the permuted index is being structured on magnetic tape for a computerized typographic printout is an important step in the evaluation of an automated thesaurus. This structure of terms appears to have some of the best possibilities for an open-end vocabulary which can be automatically updated as new technology and terms are evolved. This process involves a computerized frequency count of term usage in the series of program planning and work activity reports related to the Army's published qualitative requirements. The total concept needs more study, and will probably be the subject of a later paper. Plans are now being made to make the thesaurus one of the permanent major files of RODATA.

FIELD AND GROUP NO.	COSATI, DD FORM 1630 OR TEST TITLE	TYPE OF ACTION
01 00	AERONAUTICS	
01 01	Aerodynamics	Deleted by TEST
01 04	Aircraft Flight Instrumentation	Group simplified by DDC
05 00	BEHAVIORAL AND SOCIAL SCIENCES	
05 02	Information Sciences	Changed by TEST
05 08	Man-Machine Relations	Deleted by TEST
05 10	Psychology	Name simplified by TEST
06 00	BIOLOGICAL AND MEDICAL SCIENCES	
06 10	Industrial (Occupational) Medicine	Deleted by TEST
06 12	Medical Equipment and Supplies	Simplified by TEST
06 22	Biophysics	Added by DD Form 1630
07 00	CHEMISTRY	
07 04	Physical and General Chemistry	Expanded by TEST
07 06	Analytical Chemistry	Added by DD Form 1630
08 00	EARTH SCIENCES AND OCEANOGRAPHY	
08 14	Geomagnetism	Changed by TEST
10 00	NONPROPULSIVE ENERGY CONVERSION	Changed by TEST
11 00	MATERIALS	
11 06	Metals	Changed by TEST
11 13	Corrosion and Degradation	Added by TEST
13 00	MECHANICAL, INDUSTRIAL, CIVIL AND MARINE ENGINEERING	
13 09	Machinery, Tools, and Industrial Equipment	Changed by TEST
13 10.1	Submarine Engineering	Added by DDC
14 00	METHODS AND EQUIPMENT	
14 06	Research	Added by TEST
14 07	General Concepts	Added by TEST
14 08	Proposed to COSATI by DDC	Deleted by TEST
14 09	Geometric forms	Added by TEST
15 00	MILITARY SCIENCES	
15 02	Chemical, Biological, and Radiological Operations	Changed by TEST
15 03.1	Antimissile Defense	Added by DDC
16 00	MISSILE TECHNOLOGY	
16 04.1	Air-and-Space-Launched Missiles	Added by DDC
16 04.2	Surface-Launched Missiles	Added by DDC
16 04.3	Underwater Launched Missiles	Added by DDC
17 00	NAVIGATION, COMMUNICATIONS, DETECTION, AND COUNTERMEASURES	
17 02.1	Radio Communications	Added by DDC
17 11	Miscellaneous Detection	Added by TEST
18 00	NUCLEAR SCIENCE AND TECHNOLOGY	
18 01	Fusion Devices (Thermonuclear)	Deleted by TEST
18 05	Nuclear Power Plants	Deleted by TEST
18 09	Reactor Technology	Changed by TEST
18 12	Reactors (Power)	Deleted by TEST
18 13	Reactor (Non-Power)	Deleted by TEST
18 14	SNAP Technology	Deleted by TEST
19 00	ORDNANCE	
19 01	Ammunition, Explosives, and Pyrotechnics	Added by DDC
20 00	PHYSICS	
20 08	Particle Physics and Nuclear Reactions	Changed by TEST
20 10	Quantum Theory and Relativity	Changed by TEST
20 11	Mechanics	Changed by TEST
21 00	PROPULSION, ENGINES, AND FUELS	Changed by TEST
21 01	Air-Breathing Engines	Deleted by TEST
21 08	Rocket Engines	Changed by TEST
21 08.1	Liquid Rocket Motors	Added by DDC
21 08.2	Solid Rocket Motors	Added by DDC
21 09.1	Liquid Rocket Propellants	Added by DDC
21 09.2	Solid Rocket Propellants	Added by DDC
21 10	Engine Components	Added by TEST
21 11	General Engine Concepts	Added by TEST
21 12	General Propulsion Concepts	Added by TEST

FIGURE 3. Changes or Additions to COSATI List at Field or Group Level

KEYWORDS AND FIELD OF INTEREST TERMS	CODE FOR DD FORM 1499 (RTR)	DD FORM 354-2 CODE	DD FORM 1630 CODE	COSATI GROUPS WITH RELATED FORM 1630 TERMS
<b>ABLATIVE PLASTICS COMPOSITES</b>			11090100	
<b>ABRASIVES</b>				
Solvents, Cleaners, and Abrasives	015800		11110000	
<b>ABSORBER MATERIALS</b>				
Radar Absorber Materials			11040500	
<b>ABSORBING MATERIALS</b>				
Camouflage Absorbing Materials, Reflecting Materials, Chaff		08080302		See 1104, 1704
<b>ABSORPTANCE</b>				
Thermal Absorptance and Transmission			20131000	
<b>ABSORPTION</b>				
Absorption and Transfer of Energy in The Cell			06220100	
Radiation Absorption Studies			18060400	
<b>ABSTRACTING</b>				
Cataloging, Indexing, Abstracting			05020200	
<b>AC POWER SUPPLIES</b>				
Power Supplies - AC and DC Regulated and Unregulated		07050401		See 0905
<b>ACCELERATORS</b>				
Nuclear Accelerators			18080700	
Particle Accelerators	012200		20070000	
Reactors and Particle Accelerators		20050000		See 2007, 1812, 1813
Wind Tunnel Accelerators			10030500	
<b>ACCELEROMETERS</b>				
ACCELEROMETERS, Indicating			01040901	
<b>ACCEPTANCE</b>				
Acceptance (Load)			08080601	
<b>ACCESS</b>				
Multiple Access (Multi-Subscriber)			17020204	
<b>ACCESSORIES</b>				
Accessories (Nuclear Propulsion)		21010000		See 1805, 2106
Accessories (Propulsion Systems)		27080000		See 1310
Accessories (Propeller-Rotor Design)		01080203		See 0103
Automotive Parts and Accessories		11020000	15050501	
Components and Accessories (Electronic Equipment)		08020000		See 0901
Components and Accessories (Armored Vehicles)			19030101	
Components and Accessories (Nuclear Engines)			21060300	
Components and Accessories (Tanks)			19030902	
Computer Accessories			09020500	
Fuel Equipment and Accessories			01031800	
<b>ACCOUNTING</b>				
Accounting			05010100	
Management, Accounting and Public Relations		26020000		See 0501
<b>ACCUMULATORS (HYDRAULIC AND PNEUMATIC)</b>			13070100	
<b>ACOUSTIC</b>				
Acoustic Detection	000100	06010000	17010000	
Acoustic Intensity			20010200	
Acoustic Mines			19011301	
Acoustic Sensors		18020102	15040101	

FIGURE 4. Permuted Index Format

I have stated that the Army thesaurus is going to be produced using a computerized typographic printout (see Figure 4). These will become pages in the proposed DA Pamphlet 70-20-1, "QRI Guide to Automated Procedures." This will be a supplement to DA PAM 70-20, the "QRI Managers Guide" which is presently scheduled for publication. In addition to the thesaurus the PAM 70-20-1 will also cover formats for all files, and detailed instructions for all card, paper tape, or on-line terminal inputs and outputs for the entire RODATA System. The methods and procedures applicable to thesaurus updating will also be provided, so that in addition to regular automated review of current activity documents, each Army installation can also submit new terms for consideration. It is expected that TEST will be made into an open-ended document by that time, so that new terminology created throughout DoD will be reflected in later supplements or editions.

When once we are able to achieve a standardized method for recording new language it will be possible to produce annual or biannual revisions of the DD Form 1630. This has been a prior difficulty with classification forms such as the AFSC 220 or DD Form 558-2. Also we plan to revise the DD Form 1630 into a form suitable for direct optical reading or scanning. There may be problems with ASPR authorizations and prompt publication of revised forms for a while yet; however, there are possibilities for moving more quickly in this area. When on-line full automation is achieved, it will be possible to accelerate the entire process. If one thinks of the code for each term in the thesaurus as the equivalent of a telephone number for a person, there is no reason why the combined QRI capabilities list and ASPR bidders list operations cannot operate on a 24 hour updating routine.

#### Use of the Thesaurus

In summary, although most of them have already been mentioned in this paper, the various forms of the thesaurus will find these uses:

1. Conversion by user, civilian or Army, of DD Form 558-2 terms and coding to DD Form 1630 terms and coding.
2. Thesaurus of uniformly acceptable terms to be used as descriptors and keywords on QRI statements for use in information retrieval.
3. Meaning and use of terms appearing on 3rd and 4th tiers. Eliminate ambiguity.
4. Mechanism for adding new terms to the QRI vocabulary. Open ended arrangement that will accept new terms easily, and automatically assign them to their proper hierarchical structure and coding.
5. Preparation of registration information by new registrants. Scope notes and permuted index provide guidance to existing terminology which best describes activities and capabilities.

### Bibliography

1. Peirce, J. G. and Shannon, W. B., "Interaction Between a Computerized Retrieval System and a Commercially Available Microfilm Cartridge File for Backup Data from Diverse Locations," Proceedings of the American Documentation Institute, Volume IV, New York City, October 1967.
2. Peirce, J. G. and Segal, J. J., "Source Data Capture and Conversion - Simplicity or Complexity," Author Panel No. 4, Proceedings American Society for Information Science, Volume 5, Columbus, Ohio, Greenwood, New York City, October 1968.
3. Committee on Scientific and Technical Information, COSATI Subject Category List (DoD-Extended), Defense Documentation Center (DDC), Alexandria, Va., December 1965, AD 624 000.
4. Committee for Revision of DD Form 558-2, Minutes of 1st Meeting, US Army Electronics Command, Ft. Monmouth, N.J., May 1965.
5. TEST, Thesaurus of Engineering and Scientific Terms, DoD, Office of Naval Research, Washington, D.C., 1967, AD 672 000.

PLANS FOR UPDATING THE  
THESAURUS OF ENGINEERING AND SCIENTIFIC TERMS

---

SURVEY OF USERS

Frank Y. Speight  
Director - Information Program  
Engineers Joint Council

Purpose and Content

The intended use of the thesaurus is two fold -- one for reference for indexing and retrieval by those who do not have or use controlled vocabularies and the other, as a base for building controlled vocabularies for indexing and retrieval. The thesaurus which was developed during 1966 and 1967 and published in the summer of 1968 does not provide flexibility for adding authorized new terms for new concepts as they are developed. This can only be provided by periodic updating of the thesaurus and plans for such updating will be described. Also included in the paper are preliminary results of a survey of some 4000 users of the thesaurus. Efforts to standardize the Rules and Conventions which specify the thesaurus structure and logic will be described.

Purpose of the Thesaurus

In 1961 the American Institute of Chemical Engineers published the Chemical Engineering Thesaurus which was based upon a thesaurus used internally by the DuPont Company for indexing company reports. The ASTIA Thesaurus soon followed and in 1964, Engineers Joint Council published the Thesaurus of Engineering Terms. These thesauri then became the springboard for the development of a large number of specialized thesauri by organizations endeavoring to organize and manage the specialized literature in their fields.

On superficial examination it might appear that there were two conflicting purposes in the development of a broad thesaurus covering the whole field of technology -- one being that the appearance of such a thesaurus would make it unnecessary for organizations to develop their own and two, the appearance of such a thesaurus would facilitate the development of specialized thesauri. Actually, the thesaurus is useful in both instances with certain reservations. It should be pointed out that the structure of the thesaurus, rules out many terms that are useful in indexing and therefore, the thesaurus is not a complete indexing vocabulary. The Thesaurus Rules and Conventions which have already been discussed point out that there are two major categories of terms included in the thesaurus but that a number of terms such as the names of things and the like are not included although these are useful indexing concepts.

The two major uses of the thesaurus are described on page three of the Thesaurus of Engineering and Scientific Terms. The first -- indexing and retrieval -- applies to those who do not have or have access to a specialized thesaurus in the field of interest in which indexing and retrieval is being done and therefore, reference is made to the use of the thesaurus for this purpose. The second area, that of vocabulary building, is probably the major use of the thesaurus.

Also on page three of the thesaurus the point is made that an interdisciplinary thesaurus such as the Thesaurus of Engineering and Scientific Terms resolves a number of term conflicts where the use may be different from one field to another. This, of course, makes the thesaurus more useful in establishing the basis for compatibility of information systems in different fields.

#### Extent of Use

Engineers Joint Council has sold both in the United States and abroad about 4500 copies of the Thesaurus of Engineering and Scientific Terms and several copies of the Magnetic Tape Edition. Probably several thousand more than this number have been distributed by the Defense Documentation Center to authorized DDC users. Of the EJC distribution of the book, about a third of the total number of copies have gone overseas. Of the books distributed in the United States Table 1 gives the percentages acquired by various types of organizations.

TABLE 1 - Distribution of the Thesaurus  
in the United States

	<u>Percentage</u>
Government, Federal, State, Local	11.5
Companies	51.4
Bookstores	10.4
Universities	12.5
Individuals	3.8
Professional societies, non-profit institutions, etc.	<u>10.4</u>
	100

Engineers Joint Council continues to sell the thesaurus at a rate of about 150 per month.



### Tentative Plans for Revision

The published thesaurus became available in August 1968. It has at this writing been in use for approximately one year although the editorial content had been frozen about a year earlier, the summer of 1967. It is a well known fact that extensive world-wide research and development efforts continue to introduce new terms and concepts into the language. An ideal situation would be continuous updating of the thesaurus to accommodate these changes as they occur. This is not possible for a printed thesaurus and the next best alternative is to publish revised editions at periodic intervals. Another possibility would be periodic publishing of supplements to the book.

For any organization to assume the responsibility of keeping a thesaurus up to date, it is necessary to justify to the management of that organization that the thesaurus is serving a useful purpose commensurate with the costs incurred in preparation and publication of the thesaurus. If this responsibility can be justified, then the next question is, can this work be done on a self-supporting basis or is a subsidy required? While the economic aspects of thesaurus work are beyond the scope of this paper, nevertheless, it is an important consideration both for Engineers Joint Council and any other organization that might have or assume that it has the responsibility for leadership in thesaurus work. By mid 1970, Engineers Joint Council will have recovered all of its out-of-pocket costs in thesaurus work and will accumulate a small surplus but not enough to support continuing revision work.

### Revision Plans

A decision has been made by Engineers Joint Council that a revision of the thesaurus cannot and must not be done by the method used in the original construction of the thesaurus. A less expensive but equally satisfactory or better method must be developed. The following operations are involved in thesaurus updating regardless of the specific techniques to accomplish them.

- . Candidate term acquisition
- . Clerical and computer ordering
- . Analysis and editing by lexicographers assisted as necessary by subject specialists
- . Updating the machine readable thesaurus
- . Typesetting and printing

There are two major approaches to candidate term acquisition. One is to survey the thesaurus users and others asking them to contribute new terms or changes in terms from the thesaurus. The other method involves the analysis of large machine readable data bases covering the scope of the thesaurus and using automatic indexing techniques and other computer analysis techniques for selecting out terms that are in these machine readable data bases and comparing them by the use of concordance programs with terms in the existing thesaurus and printing out lists of terms that match and terms that don't match.

The first method would allow for correction of known deficiencies in the thesaurus and the addition of new terminology but it would suffer from a lack of uniform coverage of all the fields included in the thesaurus. The second approach -- that of computer analysis of terms -- is less certain of its outcome than the other method but assuming a broadly based data base to analyze, a more uniform coverage of the literature might be expected.

When the candidate terms have been acquired, they will be suitably coded as to the source and displayed in alphabetical sequence in relation to the existing thesaurus structure. Professional lexicographers and indexers will then be employed to analyze the candidate terms and to introduce certain revisions into the thesaurus based on this analysis. In case of doubt as to the proper meaning or appropriateness of certain terms, subject experts will be consulted probably by telephone or by informal meetings. It is not intended that any regular series of meetings such as those used in the development of the other thesaurus will be used in this instance.

No timetable for the revision effort has been established and it is anticipated that it will probably not occur before 1972 or 1973.

There is a possibility that by the time Engineers Joint Council is ready to prepare a revised edition of the thesaurus, there will be an operational indexing system or systems that are sufficiently broad such that the updating of the thesaurus can be tied to this operating system. If this could be done, then a major criticism of the thesaurus that it is not sufficiently in touch with the current literature would be eliminated and the thesaurus in effect would be continuously updated by the operating service.

#### Survey of Users

In order to get some feedback from the users of the thesaurus, EJC early in 1969 initiated a survey of those who had purchased the thesaurus and to date, (August 21), approximately 700 postcard questionnaires had been returned. The survey objective was first to send out a very simple questionnaire to everyone who had bought the thesaurus for the initial purpose of identifying those having a high interest in thesaurus use and revision participation. A second more detailed questionnaire is to be sent to those who have expressed a high degree of interest.

The following questions were asked in the initial survey:

1. Do you use the Thesaurus?  
Regularly, occasionally, not at all
2. Do you use the Thesaurus Rules and Conventions?  
As is, with modifications, don't use them
3. Would you contribute terms for a revised edition?  
Yes, No

TABLE 2

<u>Results of Evaluation</u>	<u>Questionnaire</u>
Questionnaires distributed approximately	3000
Questionnaires returned approximately	700
Use Thesaurus regularly	223
Use Thesaurus occasionally	395
Use the Rules as is	127
Use the Rules with modifications	231
Will contribute terms for revised edition	268

The results of the returns as of August 12, 1969 are given in Table 2 above.

We also asked for comments in addition to completing the questionnaire and a few have been received. Selected comments follow:

- . From an electronics concern: "We use the thesaurus for all cataloging -- books, reports, vertical file material, and filing of literature searchers. It is a very useful tool... We hope the thesaurus is now fairly stable and the new terms will be added rather than further revision of usable old terms. The additional management terms have been helpful... It facilitates cataloging to have all the terms we use in one thesaurus."
- . From an engineering company librarian: "If, instead of creating a new and useful tool, some compatible system could have been worked out with the Library of Congress, it might have been much more universally acceptable and usable. Perhaps such thought was given before you forged headlong into your separate venture... You ought to be complimented on the enormous amount of work, and in depth, which you have accomplished in your thesaurus." Congratulations! You recognized a serious need and you produced a useful working tool."
- . From Sweden: "We are cooperating through Nordforsk (Scandinavian Council for Applied Research) to get unification of the structure of a general thesaurus and branch thesauri."
- . From an individual purchaser: "Completely dissatisfied, not what I expected it to be. Not useful to me at all."

- . From an optical equipment company: "It is worthless to us and was returned."
- . From an engineering company: "I am the only person who has had an interest in the Thesaurus in an engineering - design - consulting organization of 450 persons, about 200 engineers. I have had no need or benefit from the Thesaurus."
- . From an aviation company: "It would be nice if closely related terms could have a brief definition describing the difference of meaning."
- . From a company librarian: "The present thesaurus has been most adequate."
- . From a consulting company librarian: "I am extremely impressed with the quality and depth of coverage of this thesaurus and have cataloged my home collection by COSATI as well as our company library."
- . From a research librarian: "Most excellent! Should be adopted as standard for all U.S. retrieval systems."
- . From a user in England: "I have always advised that users should stick to the Rules and Conventions of which I thoroughly approve."

We intend to follow up with a more detailed questionnaire to those who have expressed interest in cooperating. A great many more probably would have cooperated except they indicated a shortage of personnel.

#### Standardization of Rules and Conventions

Through the Engineers Joint Council representative on the USA Standards Institute Sectional Committee Z39 on library work and documentation, EJC submitted the Rules and Conventions for consideration as a possible standard several years ago. No action was taken on this proposal until mid year 1969 at which time I was asked to chair a new subcommittee of Sectional Committee Z39 to develop a draft standard Thesaurus Rules and Conventions document for consideration by USASI. I have accepted this chairmanship and intend to organize the subcommittee and work towards the development of such a standard. It is my opinion that the thesaurus itself should not be standardized but the Rules and Conventions are a good topic for standardization.

Such a standard would enable all thesaurus builders and users to establish an essentially identical format which would improve compatibility among information systems using a thesaurus for the vocabulary control mechanism.

I should welcome any constructive suggestions from ASIS members concerning the need for thesaurus revision and the means by which it might be accomplished.